CAPÍTULO 2

PLANTEAMIENTO MATEMÁTICO DEL PROBLEMA

En este capítulo se hace una descripción de los datos originales del problema y de cómo se van procesando para resolver el problema concerniente. En la sección 2.1 se describe el modelo matemático de acuerdo a los datos del problema, en la sección 2.2 se detalla como obtener el catálogo de preguntas para la solución del tema y finalmente en la sección 2.3 se aplican los conceptos de la sección 2.1 y 2.2 para obtener una matriz donde se puedan almacenar todos los datos planteados.

2.1 DESCRIPCIÓN DEL MODELO MATEMÁTICO

La lista original de los alelos HLA-I-A consta de 278 alelos agrupados en 21 subclases y con una longitud de 545 caracteres (A, C, G, T) cada alelo. A partir de esta información se plantea el modelo matemático del problema [7].

Se tiene un conjunto R de 278 alelos cada uno con 545 elementos en el conjunto $L = \{A, C, G, T\}$. Si el n-*ésimo* término $1 \leq n \leq 545$ de un alelo es una letra $X \in L$ se dice que la letra X se encuentra en la posición n del alelo que se representa. Los alelos pertenecientes al conjunto R están agrupados en 21 subclases o subconjuntos que no se intersectan entre sí.

Sin embargo, los datos originales fueron depurados. Esto es porque se eliminaron las posiciones que no dan información valiosa, es decir, las columnas en las que todos los

alelos del conjunto *R* tienen la misma letra. Esta depuración eliminó 416 posiciones, dejando un total de 129 posiciones. No obstante es importante comentar que se debe mantener la posición original, es decir, la numeración de las 129 posiciones no es consecutiva.

Es importante señalar que para simplificar la notación se reserva el término "clase" para referirse a una subclase del conjunto de alelos HLA-I-A porque más adelante se introducirá el concepto de "biclase" para representar los pares de alelos de cada individuo.

De esta manera los datos originales constan de un conjunto que se denota por la letra *S*, cuyos elementos son 278 alelos con 129 posiciones, agrupados en 21 subclases. Además, el conjunto de clases forma una partición *P* de *S*, es decir, que *P* es una familia de 21 subconjuntos disjuntos de S denotados por números, con $P = \{01,02,03,11,23,24,25,26,29,30,31,32,33,34,36,43,66,68,69,74,80\}$.

Al ser todos los alelos de la misma longitud (129) se puede construir a partir de S una matriz de la cual las filas corresponden a los alelos y las columnas a las posiciones. De esta manera se tiene una matriz *M* de 278 * 129, con entradas del conjunto $L = \{A, C, G, T\}$. A continuación se presenta un extracto pequeño de esta matriz.

Tabla 2.1 Datos originales resumidos

	05	08	17	19	24	25	29	33	40
A*010101	C	C	G	A	A	T	A	G	G
A*010102	C	C	G	A	A	T	A	G	G
A*0102	C	C	G	A	A	T	A	G	G
A*020101	T	A	T	G	T	T	A	C	G
A*020102	C	A	T	G	T	T	A	C	G
A*020103	T	A	T	G	T	T	A	C	G
A*020104	T	A	C	A	T	T	A	C	G
A*020105	T	A	C	A	T	T	A	C	G
A*03010101	C	T	G	T	C	A	C	G	T
A*03010102N	C	T	G	T	C	A	C	G	T
A*4301	C	C	G	A	T	A	C	G	G
A	0	5	0	6	3	3	8	0	0
C	7	4	2	1	2	0	3	5	0
G	0	0	6	3	0	0	0	6	9
T	4	2	3	1	6	8	0	0	2

En la Tabla 2.1 se puede observar un conjunto de 11 alelos (A*010101, A*010102,..., A*4301) agrupados en 4 clases (01, 02, 03, 43). La letra A que aparece antes del número de cada alelo, indica el grupo con el que se está trabajando. Los siguientes dos dígitos determinan el número de la clase a la que pertenece cada alelo, así se puede observar que la clase 01 tiene 3 alelos, la clase 02 cuenta con 5 alelos, la clase 03 contiene 2 alelos y finalmente la clase 43 tiene 1 alelo. Los dígitos restantes representan el número de alelo dentro de cada clase, pero esta información no es relevante para el problema.

Las primeras líneas de esta tabla, indican el número de la posición. Así el alelo A*010101 tiene la letra C en la posición 5 y 8, la letra G en la posición 17, etc. Se debe notar que el número de las posiciones no es consecutivo. Esto se debe a la depuración de los datos, en donde se eliminaron las columnas o posiciones donde se presentaba la misma

letra para todos los alelos del conjunto *R*. Finalmente, en la parte inferior de la tabla se enlistan las cuatro letras del conjunto *L* y en cada columna la frecuencia con que aparecen en la posición correspondiente. De la misma forma se presentan los datos originales del conjunto *S* contando con 278 alelos y 21 clases con distintos números de elementos en cada una, desde un solo alelo en algunas clases hasta 73 alelos en la clase más grande. Los datos originales se pueden consultar en el apéndice E, el cual consta de un disco compacto donde se archivarán los resultados obtenidos en esta tesis.

Una vez descrito el modelo matemático para los datos originales del problema, ahora se va a desarrollar para trabajar con el problema de pares de alelos. En esta tesis se genera una matriz de "biclases" M_b, combinando cada uno de los elementos del conjunto *P*, realizando combinaciones de dos alelos sin repetición entre las 21 clases, es decir, aparece la combinación de las clases 01x02 pero no es necesario la combinación 02x01 y así se genera el conjunto:

$$P\text{`} = \{01 \times 01, 01 \times 02, 01 \times 03, ..., 01 \times 80, 02 \times 02, 02 \times 03, ..., 02 \times 80, ... 80 \times 80\}$$

Par obtener el número de bialelos que debe contener la matriz M_b, es necesario aplicar la expresión $\frac{n(n+1)}{2}$, donde n = 278 alelos por lo tanto se tienen 38781 bialelos agrupados en 231 biclases. Para obtener esta matriz M_b de 38781x129 se realizó un programa computacional en lenguaje C, el código del programa se presenta en el apéndice A y el archivo donde se encuentra la matriz de bialelos se anexa en el apéndice E (disco compacto). Para tener una idea más clara de esta matriz se presenta un pequeño ejemplo

Tabla 2.2 Matriz de bialelos

	05	08	17	19	24	25	29	33	40
A*01	C	C	G	A	A	T	A	G	G
A*01	C	C	G	A	A	T	A	G	G
A*01	C	C	G	A	A	T	A	G	G
A*01	C	C	G	A	A	T	A	G	G
A*01	C	C	G	A	A	T	A	G	G
A*01	C	C	G	A	A	T	A	G	G
A*02	T	A	T	G	T	T	A	C	G
A*01	C	C	G	A	A	T	A	G	G
A*02	C	A	T	G	T	T	A	C	G
A*01	C	C	G	A	A	T	A	G	G
A*02	T	A	T	G	T	T	A	C	G
A*01	C	C	G	A	A	T	A	G	G
A*02	T	A	C	A	T	T	A	C	G
A*01	C	C	G	A	A	T	A	G	G
A*02	T	A	C	A	T	T	A	C	G
A*01	C	C	G	A	A	T	A	G	G
A*03	C	T	G	T	C	A	C	G	T
A*01	C	C	G	A	A	T	A	G	G
A*03	C	T	G	T	C	A	C	G	T
A*01	C	C	G	A	A	T	A	G	G
A*43	C	C	G	A	T	A	C	G	G

En la Tabla 2.2 se puede apreciar un conjunto de 11 bialelos, los primeros 3 pertenecen a la biclase 01x01, la biclase 01x02 tiene 5 bialelos, la biclase 01x03 cuenta con 2 bialelos y la biclase 01x43 con 1 bialelo para este ejemplo. También se puede notar que los dígitos que representaban el número de alelo dentro de la clase se eliminaron porque para el problema solo interesa saber a que clase pertenece cada alelo y específicamente a que biclase pertenece los bialelos de una persona.

Ahora se desea analizar la información obtenida en la matriz M_b y transformarla para poder aplicar un método heurístico que nos pueda dar una solución inicial y esto se describe en el siguiente apartado.

2.2 ETAPA DE PREGUNTAS

Como ya se explicó anteriormente el problema se puede resolver haciendo preguntas al código genético y así obtener un conjunto de preguntas paralelas que permitan establecer la biclase a las que pertenecen el par de alelos HLA de un ser humano. También se expresó que mediante el uso de reactivos específicos es posible determinar que letra existe en una posición definida y que utilizando un solo reactivo se puede determinar las letras existentes en 20 posiciones consecutivas del código. Y es precisamente este tipo de preguntas en 20 posiciones las que se utilizan para desarrollar la heurística de este trabajo.

Es decir, utilizando la información proporcionada en la matriz *M* de 278x129 donde las filas corresponden a los alelos y las columnas a las posiciones pero se debe recordar que el número de las posiciones no es consecutivo (porque fueron eliminadas las posiciones que tenían la misma letra para todos los alelos). Partiendo de esta información se genera un catálogo de preguntas considerando 20 posiciones enumeradas consecutivamente. Explicando esto con más detalle y utilizando la información de la Tabla 2.1 el primer conjunto de 20 posiciones serían 5, 8, 17, 19 y 24 aunque no son números consecutivos se deben consideras los números de las columnas depuradas (5, 6, 7, 8,…, 20, 21, 22, 23, 24). Entonces se requiere saber cuales serían las combinaciones posibles de letras para los 11 bialelos en estas posiciones, de esta manera las preguntas obtenidas serían 6, como se muestra en la Tabla 2.3.

Tabla 2.3 Preguntas para las posiciones 5, 8, 17, 19 y 24

Pregunta	5	8	17	19	24
1	C	C	G	A	A
2	C	A	T	G	T
3	C	T	G	T	C
4	C	C	G	A	T
5	T	A	T	G	T
6	T	A	C	A	T

Ahora bien, continuando con el ejemplo de la Tabla 2.1, los conjuntos de 20 posiciones enumeradas serían:

5	8	17	19	24	
8	17	19	24	25	
17	19	24	25	29	33
19	24	25	29	33	
24	25	29	33	40	

Y por cada conjunto de 20 posiciones se obtienen todas las combinaciones posibles de letras y de esta forma para la Tabla 2.1 se obtiene un catálogo de 25 preguntas.

Tabla 2.4 Catálogo de preguntas para los datos originales resumidos

Preguntas	Combinación de 20 posiciones	Combinación de letras
1	5, 8, 17, 19, 24	C C G A A
2	5, 8, 17, 19, 24	C A T G T
3	5, 8, 17, 19, 24	C T G T C
4	5, 8, 17, 19, 24	C C G A T
5	5, 8, 17, 19, 24	T A T G T
6	5, 8, 17, 19, 24	T A C A T
7	8, 17, 19, 24, 25	C G A A T
8	8, 17, 19, 24, 25	A T G T T
9	8, 17, 19, 24, 25	A C A T T
10	8, 17, 19, 24, 25	T G T C A
11	8, 17, 19, 24, 25	C G A T A
12	17, 19, 24, 25, 29, 33	G A A T A G
13	17, 19, 24, 25, 29, 33	T G T T A C
14	17, 19, 24, 25, 29, 33	C A T T A C
15	17, 19, 24, 25, 29, 33	G T C A C G
16	17, 19, 24, 25, 29, 33	G A T A C G
17	19, 24, 25, 29, 33	A A T A G
18	19, 24, 25, 29, 33	G T T A C
19	19, 24, 25, 29, 33	A T T A C
20	19, 24, 25, 29, 33	T C A C G
21	19, 24, 25, 29, 33	A T A C G
22	24, 25, 29, 33, 40	A T A G G
23	24, 25, 29, 33, 40	T T A C G
24	24, 25, 29, 33, 40	C A C G T
25	24, 25, 29, 33, 40	T A C G G

Para poder obtener todas las preguntas posibles para la matriz *M* de 278x129, se realizó un programa en leguaje C, el código de este programa se presenta en el apéndice B, se obtuvo un catálogo de 1117 preguntas $E = \{A_1, A_{2,} A_3, ..., A_{1117}\}$, el cual se anexa en el apéndice E. Este conjunto son todas las posibles preguntas que se pueden realizar para encontrar un alelo específico, sin embargo el interés de esta tesis es obtener un catálogo mínimo de preguntas para identificar un bialelo determinado; mientras tanto este conjunto *E* es utilizado para construir una "matriz M_E de evaluación" que se describe a continuación.

2.3 MATRIZ DE EVALUACIÓN

Cuando se realizan las preguntas al código genético de que si existe una o varias letras específicas en una o varias posiciones, en este caso en particular en 20 posiciones consecutivas, la respuesta que se puede obtener es Sí ó No. Pero como ya se explicó con anterioridad, si la respuesta es afirmativa no se sabe si es porque uno o los dos pares de alelos presentan la letra por lo tanto se tiene un perdida de información.

Utilizando el conjunto E de 1117 preguntas, a cada bialelo de la matriz de biclases M_b se efectúa cada una de estas preguntas, de esta manera, si uno de los alelos o los dos presenta la combinación de letras de la pregunta E_x, la respuesta es afirmativa y se identifica con el digito 1, si ninguno de los alelos tiene la combinación de letras de dicha pregunta se identifica con el digito 0. De esta forma se puede obtener la matriz M_E de 38781x1117 donde las filas corresponden a los bialelos y las columnas a las preguntas. Para crear esta matriz de evaluación M_E se utilizó nuevamente la programación en lenguaje C y el código el programa se agrega en el apéndice C y el archivo denominado evaluación el cual contiene a esta matriz, se anexa en el apéndice E.

A continuación se presenta un pequeño ejemplo de esta matriz y para comprender como se obtuvo esta tabla, se debe mencionar que es la combinación de las tablas 2.2 y 2.4 de este capítulo.

Tabla 2.5 Ejemplo de la matriz de evaluación

Bialelos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1*1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
1*1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
1*1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
1*2	1	0	0	0	1	0	1	1	0	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	0
1*2	1	1	0	0	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	0
1*2	1	0	0	0	1	0	1	1	0	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	0
1*2	1	0	0	0	0	1	1	0	1	0	0	1	0	1	0	0	1	0	1	0	0	1	1	0	0
1*2	1	0	0	0	0	1	1	0	1	0	0	1	0	1	0	0	1	0	1	0	0	1	1	0	0
1*3	1	0	1	0	0	0	1	0	0	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0
1*3	1	0	0	0	0	0	1	0	0	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0
1*43	1	0	0	1	0	0	1	0	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	0	1

En la Tabla 2.5 las filas corresponden a los bialelos (ver Tabla 2.2) y las columnas son las preguntas realizadas a cada bialelo (ver Tabla 2.4), se puede notar que la notación de cada bialelo se simplificó a la expresión 1*1, 1*2, 1*3, etcétera (en lugar de A01*A01, A01*A02), esto es porque solo interesa saber cuantos bialelos tiene cada biclase. De esta manera para el ejemplo de la Tabla 2.5, la biclase 1*1 cuenta con 3 bialelos, la biclase 1*2 tiene 5 bialelos, la biclase 1*3 incluye 2 y finalmente la biclase 1*43 solo cuenta con 1 bialelo.

Después de haber descrito el análisis de los datos originales y como se transformaron hasta obtener la matriz de evaluación M_E, esta matriz es la base para aplicar la entropía de Shannon y la Ganancia de la Información con los cuales se pretende encontrar una solución inicial al problema de clasificación de bialelos HLA. En la siguiente sección se hace una descripción de estos temas.